Mx modeling of methylation data:

twin correlations [means, SD, correlation]
ACE / ADE latent factor model
regression [sex and age]
genetic association analysis [SNP]

Dorret Boomsma, Nick Martin, Irene Rebollo

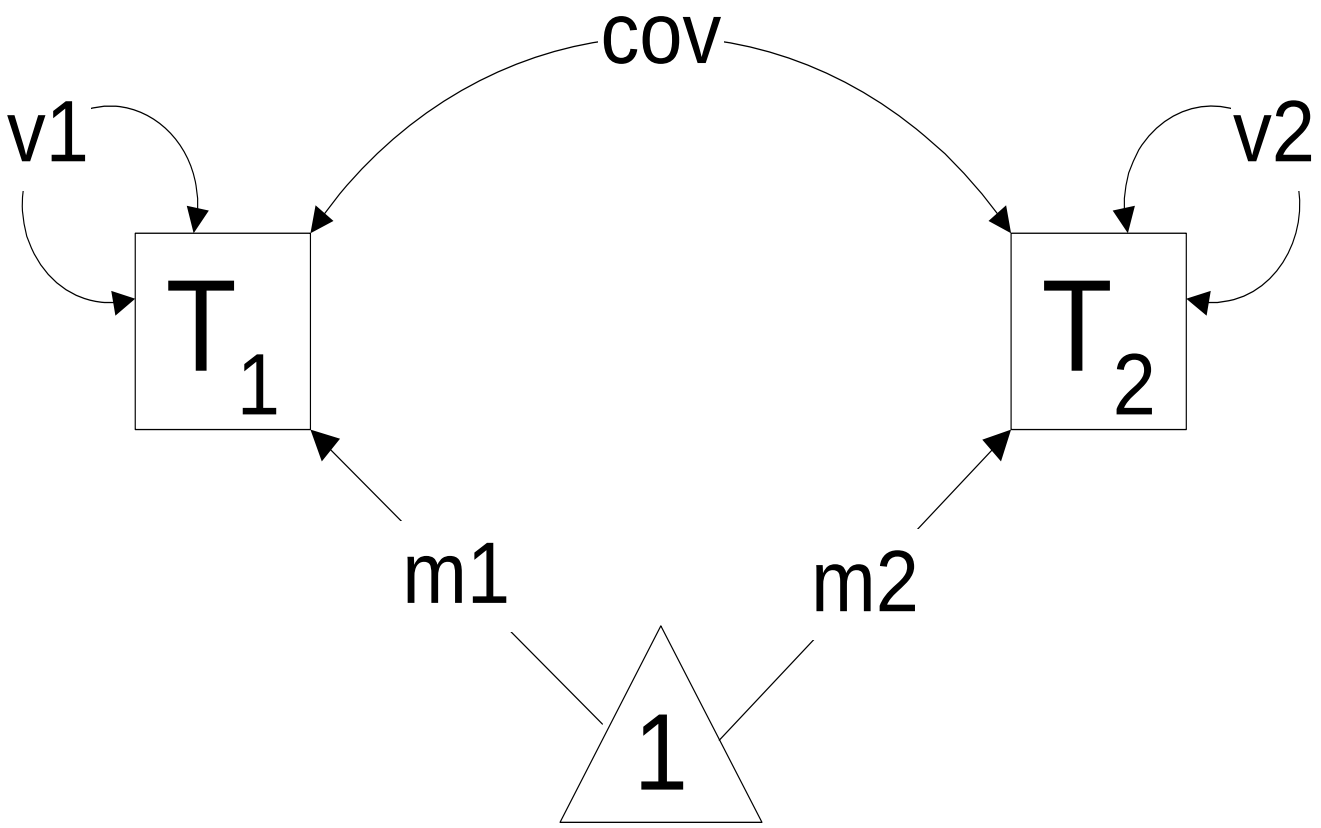
Leuven 2008

Heijmans BT, Kremer D, Tobi EW, Boomsma DI, Slagboom PE. Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. Hum Mol Genet. 2007 16(5):547-54.

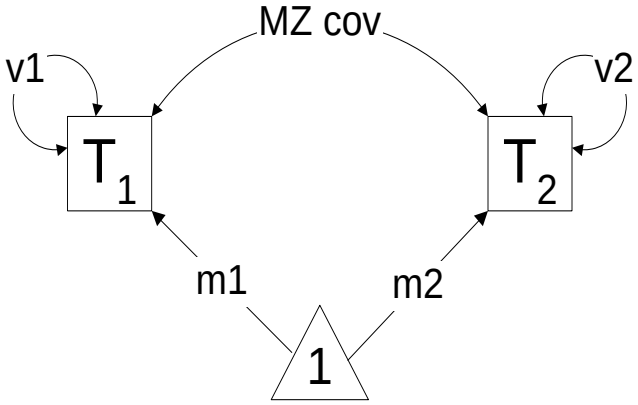
This session

- Obtain MZ and DZ twin correlations; obtain ML estimates of means and variances
- Estimate heritability from ACE model
- Extension 1: regression analysis: how well do sex and age predict methylation?
- Extension 2: regression analysis: how well do SNPs in IGF2 predict methylation? (i.e. *genetic association* test)

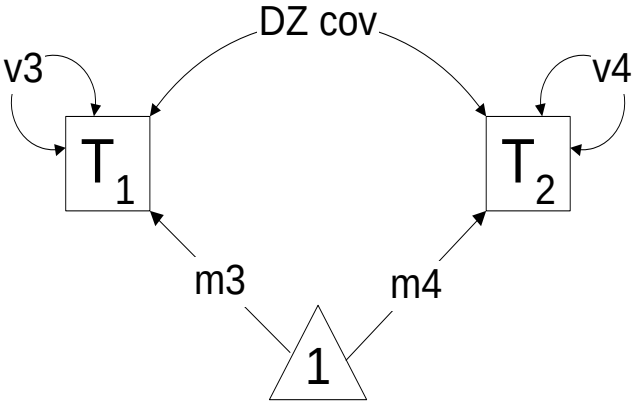
Correlation (covariance) Model



Correlation Model MZ & DZ



MZ twins



DZ twins

3 x 2 parameters: 1 mean, 1 variance, 1 covariance for MZ and DZ

(also possible to estimate separate means/variances for first- and second-born twins)

MX

- Mx script can be divided into several “groups”
- Parameters are estimated in matrices
- Matrices are defined in groups (locally)
- Matrices (or matrix elements) can be equated across groups

- To estimates MZ and DZ correlations I used 3 groups: a data definition group and 2 data groups

First group

- G1: calculation group
- Data Calculation NGroups=3
- Begin matrices;
- X dia 2 2 Free ! (standard deviation) MZ
- Y stand 2 2 Free ! correlation MZ
- S dia 2 2 Free ! (standard deviation) DZ
- T stand 2 2 Free ! correlation DZ
- G Full 1 1 free ! grand mean phenotypes MZ
- H Full 1 1 free ! grand mean phenotypes DZ
- End matrices;

Script: correlatiejob igf2_mp2_2008.mx
Data: mx_igf2_aug08_mp2.dat

Model (matrix notation)

- Covariances X*Y*X' ; ! model for MZs

- Covariances S*T*S'; ! model for DZs

```
X dia 2 2 Free          !  (standard deviation) MZ
Y stand 2 2 Free        !  correlation MZ
S dia 2 2 Free          !  (standard deviation) DZ
T stand 2 2 Free        !  correlation DZ
```

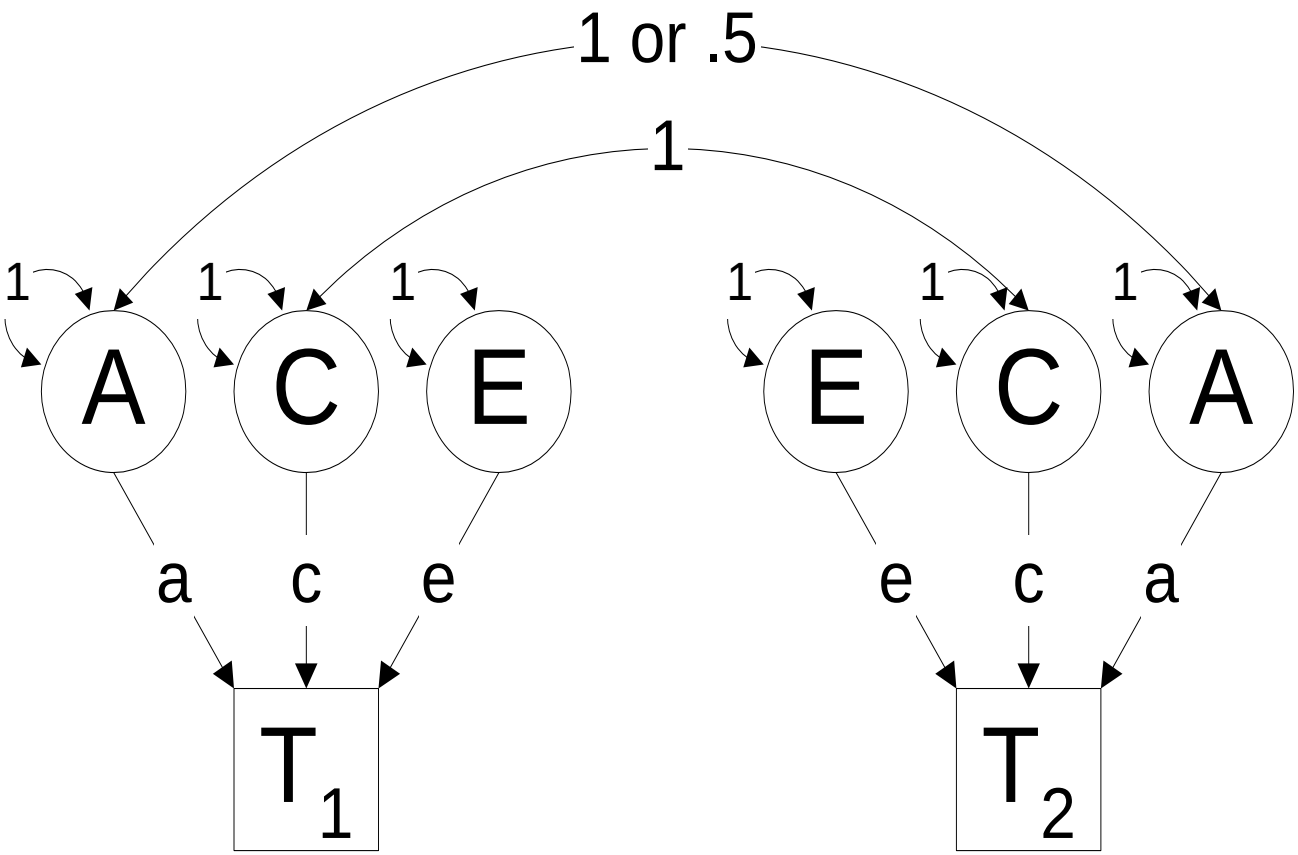
MX Output:
MZ, DZ means
MZ, DZ SDs
MZ, DZ correlations

	Mean	SD	Corr
• **MZ**	**3.04**	**0.616**	**0.66**
• **DZ**	**2.90**	**0.892**	**0.27**

6 parameters estimated; -2 * log-likelihood of data = 449.048

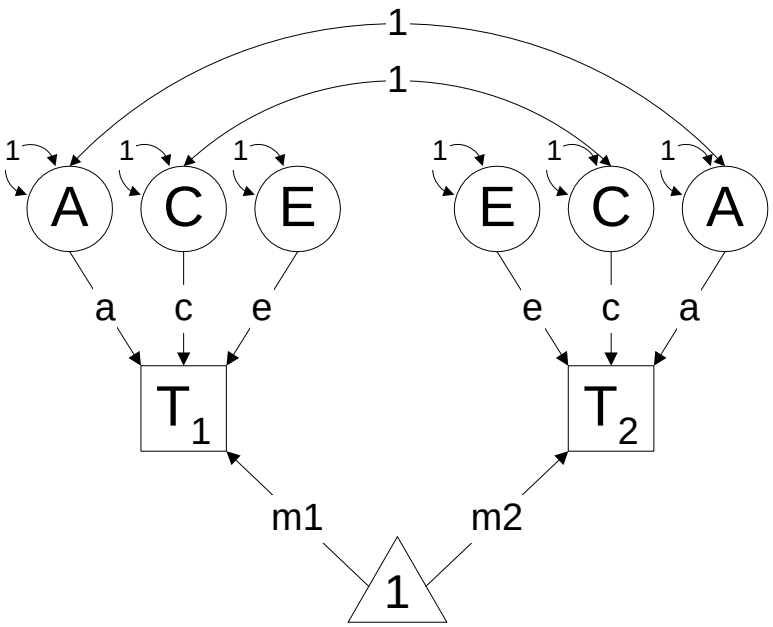
What is the heritability of this trait?

ACE Model, based on MZ & DZ data

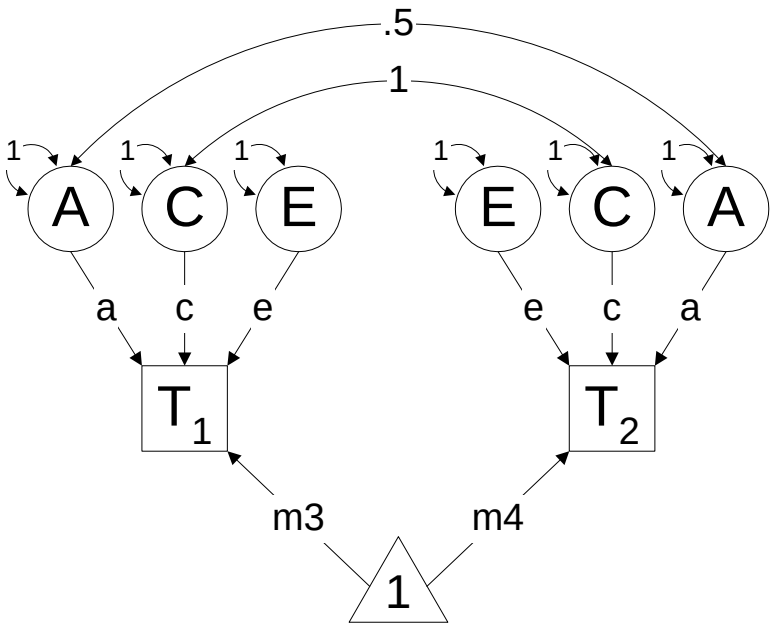


Square = observed phenotype; circle = latent variable (standardized);
a, c, e are factor loadings; if fitted to raw data model also includes means

ACE Model (+ Means)



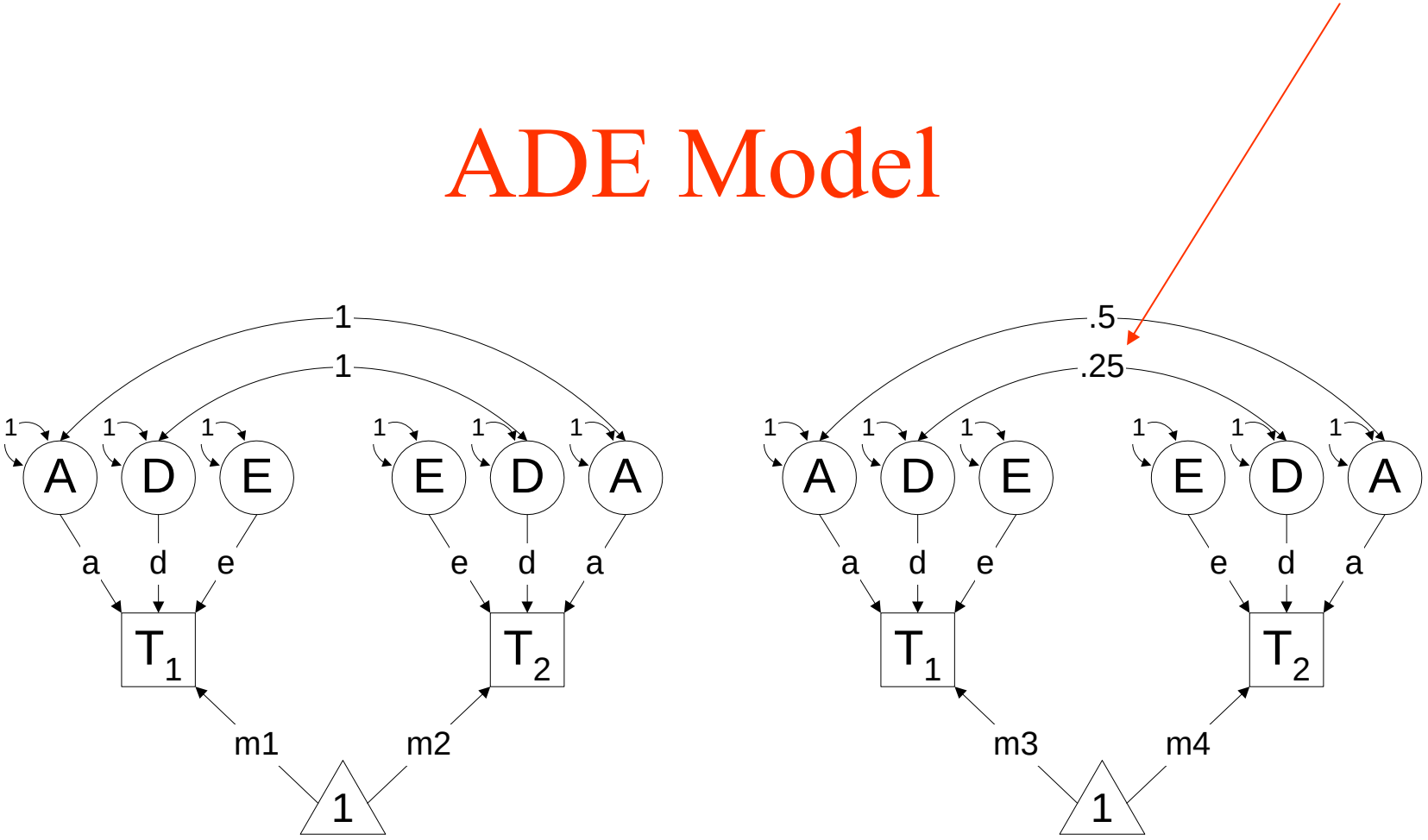
MZ twins



DZ twins

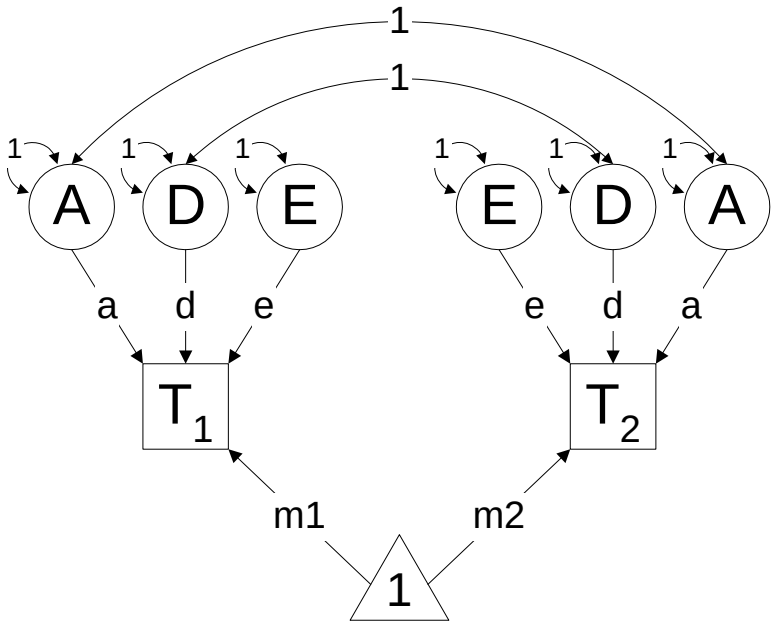
parameters: means, 3 path coefficients: a, c, e

ADE Model



When DZ correlations are much lower than MZ correlations; common environment is unlikely and genetic non-additivity (genetic dominance) may explain the data better

Individual Model & Variation



T1= aA1 + cC1 + eE1

T2= aA2 + cC2 + eE2

Var(T1) = a*a*Var(A1) + c*c*var(C1) + e*e*Var(E1) = $a^2+c^2+e^2$

Covar (T1,T2) = a*a*covar(A1,A2) + c*c*var(C1,C2) = $\alpha a^2+\gamma c^2$

Fit ACE model (and submodels)

- Modify the correlation script
- Or: take **ACE igf2_mp2_2008.mx**; fits ACE, AE, CE and E models to the data

Saturated and ACE

- Correlation script: -2LL = 449.05 (df = 6)
- ACE model: -2LL = 462.05 (df = 4)

- Does the ACE model fit worse??
- Why???

Back to correlation model

- Correlation model: -2LL = 449.05 (df = 6)
- Correlation model with equal means and variances for MZ & DZ: -2 LL = 460.651 (df = 4)
- ACE model: -2LL = 462.05 (df = 4)

Mean	**SD**	**MZ / DZ Corr**
2.964	**0.80**	**0.78 / 0.26**

- **Heritability from ACE model = 0.78%**

- CE (-2LL = 476.96) or E model (-2LL = 494.79) do not fit

Means testing, regression analysis etc. for clustered data (e.g. from twins or sibs)

- If Ss are unrelated any statistical package can be used for regression analysis, tests of mean differences, estimation of variance.
- If data come from related Ss (e.g. twins) we need to model the covariance structure between Ss to obtain the correct answer.

MX

- Mx allows us to model means and covariance structures (for dependent variables)
- If means are modeled input must be “raw data” (Full Information Maximum Likelihood - FIML)
- In addition, the user can specify “definition variables” (these are the predictors in a regression equation (= independent variables))
- The independent variables are not modeled in the covariance matrix

The likelihood of the i^{th} family

$$(2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}_i-\mu_i)'\Sigma^{-1}(\mathbf{x}_i-\mu_i)\right)$$

- **x** is vector of observed values (dependent variables) for twin1, twin2 etc
- **μ** is vector of expected values, given observed independent variables (predictors, regressors) such as age, sex, genotype etc.
- **Σ** is the variance covariance matrix of residual values after the regression effects on the expected values have been removed

Testing assumptions (1)

- Are means same for MZ & DZ?
- Are SD the same for MZ & DZ?
- *Are means same for men and women?*
- *Is there an effect of age?*

Using definition variables

- Dependent variables (phenotypes) may have missing values (e.g. -9.00)
- Independent (definition) variables are NOT allowed to have missing values (if they do, all data for that case (twin pair) are removed)

Individual differences

- Our ultimate goal is to be able to measure all causal variables so the residual variance approaches zero – except for measurement error.

- Until that time we have to continue to model variance components in terms of A, C – and E (latent (=unmeasured) constructs).

- However, if causal variables are also influenced by genes, we want to use multivariate modeling (and not correct the dependent variable)

"Means model" – or preferably model for expected individual values

- $X_i = M + B*P_i + e_i$
  - M = grand mean
  - B = regression
  - P = predictor(s), e.g. age and sex
  - e = residual term
- i stands for individual (M and B are invariant over individuals)
- read in the predictor variables in Mx

Importance of getting the means model right

- Age regression can look like C in twin model (twins are of the same age)
- If pooling data from 2 sexes, sex differences in means can create C
- Model grand mean (female) + male deviation
- Correcting for age, sex effects on means does not mean that residual variance components are necessarily homogeneous between groups – need GxE modeling (later in the week)

Definition variables: age and sex

Definition variables can**not** be missing, even if dependent variable is missing in FIML

- if dependent variable is missing, supply a valid dummy value (doesn't matter which value, as long as it is not the same as the missing code for the dependent variable!)
- if dependent variable is not missing, supply e.g. the population mean for the definition variable, or the co-twin's value – i.e. impute with care! Or delete data of this person

Mx script for age/sex correction

- Script = **Covar correlatiejob igf2_mp2_2008.mx**
- Data file = **mx_igf2_aug08_mp2.dat**

- *Or modify your old script*

- Dependent variable is a quantitative methylation score
- Data were collected on adolescent twins
- Definition variables: age and sex

- **Are effects of age and sex significant?**

Saving residuals

- Mx allows you to save residuals after baseline run and then use these as input variables for batch runs
- Option saveres

Including genotypes in the means model

- Allelic model for SNPs (2 alleles), or genotypic model (3 genotypes (0,1,2 alleles))
- For microsatellites with k alleles, k-1 deviations, k(k-1)/2 – 1 deviations!
- Missing genotypes? → not allowed

Including genotypes in the means model

- Phenotype = methylation scores

- Predictors: 1 SNP
- Coding: 0,1,2 (N of alleles)

- Script: **SNP correlatiejob igf2_mp2_2008.mx**
- Data: ***mx_igf2_aug08_mp2_noMis.dat***

- OR: modify one of the existing scripts

NB slightly different dataset

Including genotypes in the means model

- Is there evidence that this SNP has a main effect?
- No